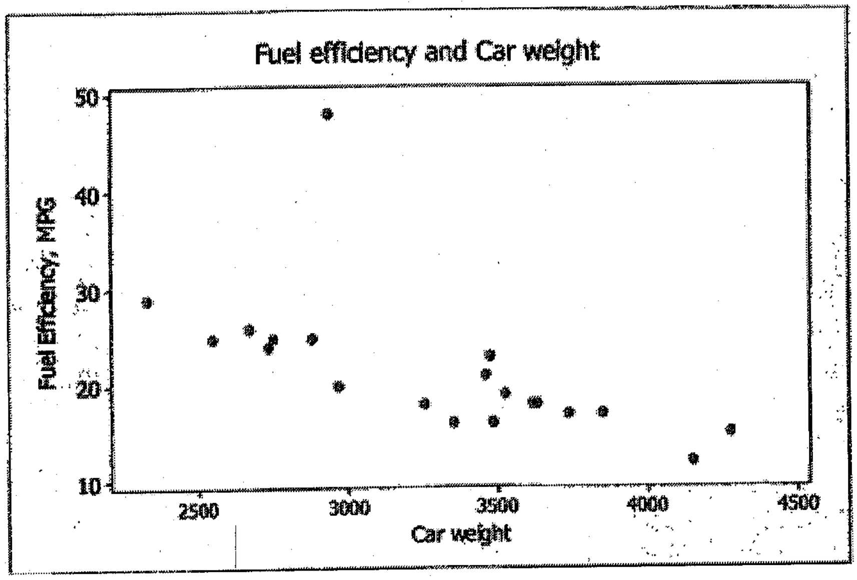
AP Statistics Ch 7,8,9 - Required Practice **Name:** Solutions **Per:** ____

The scatterplot below shows the fuel efficiency (in miles per gallon) and weight (in pounds) of twenty 2009 subcompact cars.



#1. Is there a clear explanatory variable and response variable in this setting? If so, tell which is which. If not, explain why not.

Explanatory : car weight

Response : fuel efficiency

Changes to a car's weight could cause the fuel efficiency to change, but changing a car's fuel efficiency (for example, abrupt starts) won't cause a change in car weight.

#2. Describe the association between fuel efficiency and car weight.

There is a linear, fairly strong, negative association between fuel efficiency and car weight.

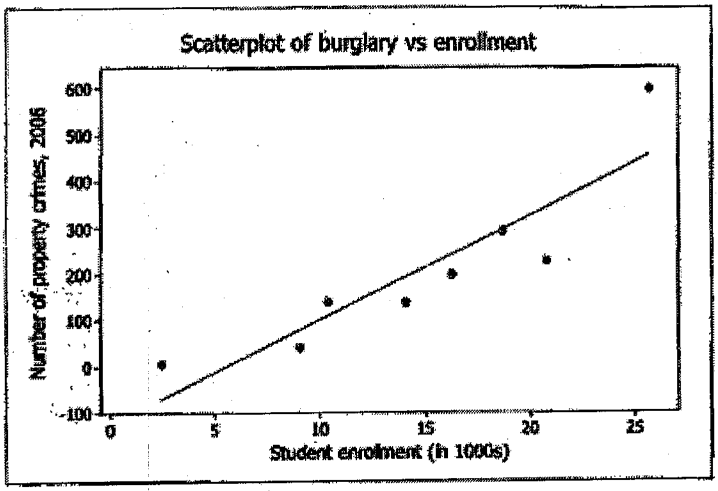
#3. Which of the following is closest to the correlation between car weight and fuel efficiency for these 20 vehicles? Explain.

[✓ r = -0.9] r = -0.6 r = 0 r = 0.4

The association is definitely negative and fairly strong.

The table and scatterplot below show the relationship between student enrollment (in thousands) and total number of property crimes (burglary and theft) in 2006 for eight colleges and universities in a certain U.S. state.

Enrollment (in 1000s) (x)	No. of Property Crimes (y)
16	201
2	6
9	42
10	141
14	138
26	601
21	230
19	294



#4. Use a calculator to find the equation of the least-squares regression line (LSRL).

$\hat{y} = -112.6 + 21.826x$

x: student enrollment (in 1000s)
y: number property crimes

#5. Interpret the slope of the LSRL in the context of the problem.

For every additional one (thousand) students enrolled, 21.83 more property crimes occur, on average.

#6. How many crimes would you predict on a campus with enrollment of 14 thousand students? (Show your work)

$\hat{y} = -112.58 + 21.83(14) = 192.62$ $\boxed{\approx 193 \text{ crimes}}$

#7. Find the residual for the campus with 14 thousand students and 138 property crimes. (Show your work). Interpret the value of the residual in the context of the problem.

predicted crimes = 193
actual crimes = 138
residual = actual − predicted
= 138 − 193
= $\boxed{-55 \text{ crimes}}$

55 fewer crimes actually occurred than the model predicts for a campus with 14 thousand students.

#8. The value of r^2 for these data is 0.801. Interpret this value in the context of the problem.

About 80% of the variation in number of property crimes is explained by the LSRL model relating crimes to student enrollment.

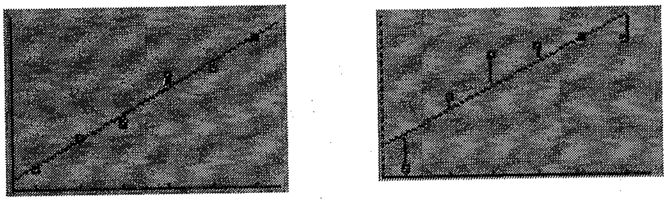
Residual Plots

A **residual plot** for a given linear regression shows the residual (r) vs. x.

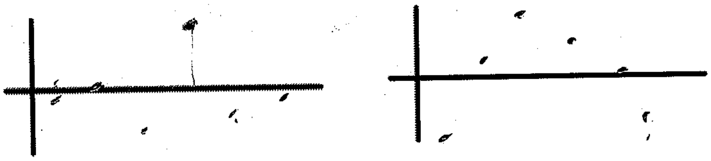
Examples: Sketch residual plots by hand for each data set

x	y
1	8
2	22
3	28
4	48
5	51
6	64

x	y
1	2
2	18
3	28
4	30
5	32
6	31



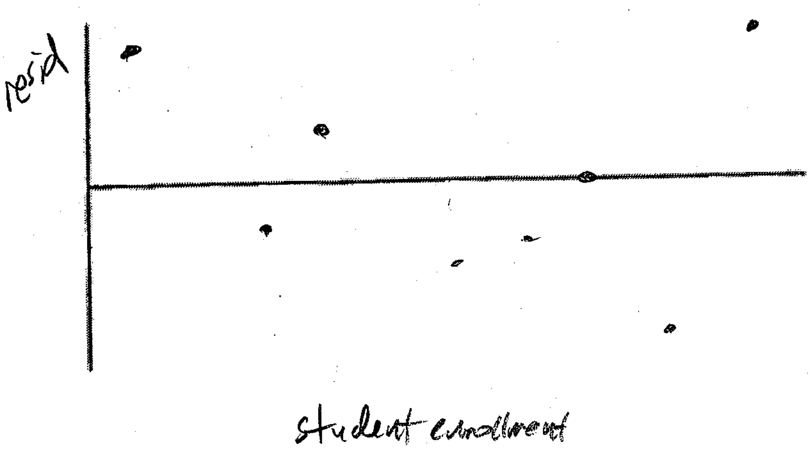
Scatter plots

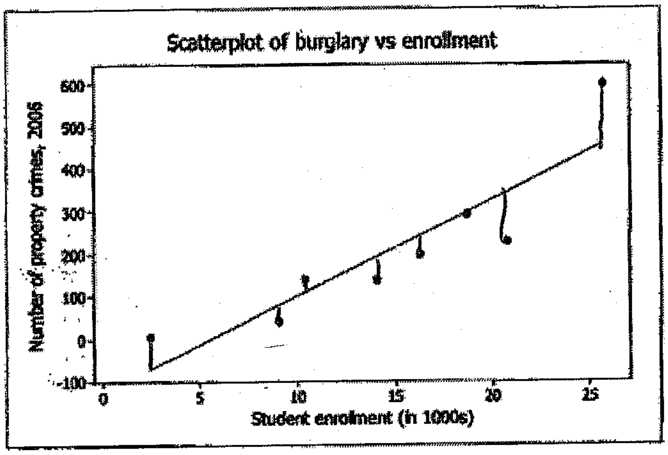


Residual plots

If the residuals are randomly scattered around '0' then you know that a linear model is appropriate. (Residuals make it easier to see non-linearity compared to scatterplots.)

#9. Use the scatterplot to make a rough sketch of the residual plot for these data.

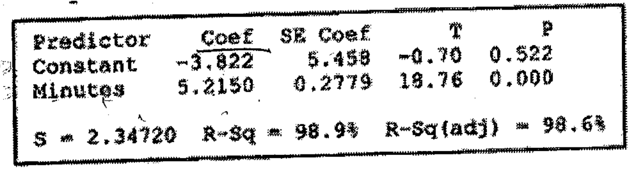




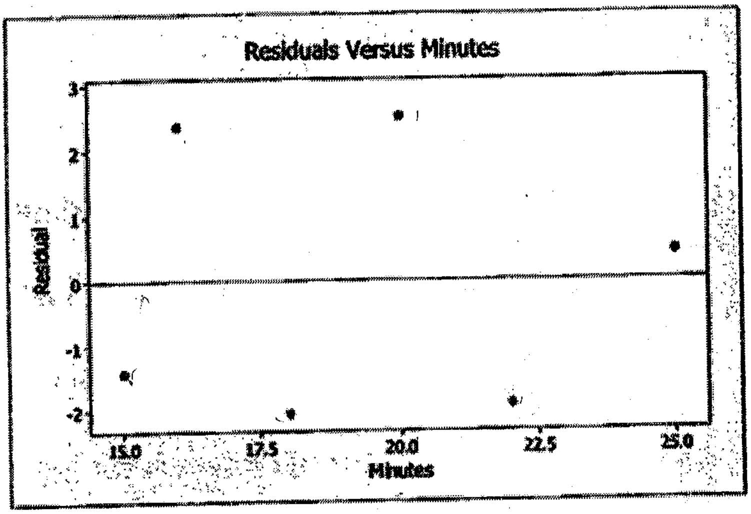
#10. Is the LSRL a good model for these data? Support your answer with appropriate evidence from your answers above.

yes, because there is no pattern in the residuals.
(Also, $r^2 = .80$ means this model is predicting much of the property crime variation.)

Exercise machine - Alana's favorite exercise machine is a stair climber. On the "random" setting, it changes speeds at regular intervals, so the total number of simulated "floors" she climbs varies from session to session. She also exercises for different lengths of time each session. She decides to explore the relationship between the number of minutes she works out on the stair climber and the number of floors it tells her that she's climbed. She records minutes of climbing time and number of floors climbed for six exercise sessions. Computer output and a residual plot from a linear regression analysis of the data are shown below.



```
Predictor    Coef   SE Coef       T      P
Constant   -3.822     5.458   -0.70  0.522
Minutes    5.2150    0.2779   18.76  0.000

S = 2.34720   R-Sq = 98.9%   R-Sq(adj) = 98.6%
```



#11. What is the equation of the LSRL? (Be sure to define any variables you use.)

$\hat{y} = -3{,}822 + 5{,}2150X$ X ; exercise time (min)
y ; number of floors

#12. Is a line an appropriate model for these data? Justify your answer.

yes, because the residual plot has no pattern.

#13. Interpret the value of the slope of the LSRL in the context of this problem.

For every additional minute of exercise, the machine reports 5,215 additional floors climbed, on average.

#14. Interpret the value of r^2 for these data in the context of this problem.

About 98,9% of the variation in number of floors is explained by the LSRL model which relates floors to exercise time.

#15. What is the correlation coefficient for this data?

$r^2 = ,989$ so $r = \pm\sqrt{,989} = \boxed{,9945}$ (positive, because slope is positive)

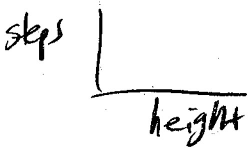
#16. Interpret the value of *s* (S=2.3472) in the context of this problem.

The difference between actual number of floors and predicted number of floors (for a given number of minutes exercise) is 2,3472 floors, on average.

1) Given summary statistics (but no data) find the slope and LSRL

$\overline{height} = 65.8333\ in$ $\overline{steps} = 16.1389$ $r = -.8734$

$s_{height} = 4.9598\ in$ $s_{steps} = 3.2527$



a) Use formula to calculate b.

b) Find y-intercept a - the centroid ($\overline{x}, \overline{y}$) will *always* lie on the LSRL.

c) Write out the LSRL.

a) $b = r\frac{S_y}{S_x}$

$b = (-.8734)\frac{(3.2527)}{(4.9598)}$

$b = -.5728\ \frac{\text{steps}}{\text{in}}$

b) $\hat{y} = \boxed{a} - .5728x$

$(\overset{\overline{x}}{65.8333}, \overset{\overline{y}}{16.1389})$ is on LSRL

$(16.1389) = a - .5728(65.8333)$

$a = 53.8482$

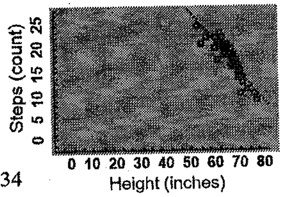
c) $\boxed{\hat{y} = 53.8482 - .5728x \quad x: \text{height (in)} \quad y: \text{steps}}$

2) Find LSRL if x and y are swapped

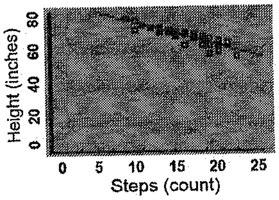
$\widehat{steps} = 53.846 - 0.5728\ height$

$\overline{steps} = 16.1389$

$r = -.8734$



$\overline{height} = 65.8333\ in$



Does r change? **no (r measures strength of association, swap does not affect r)**

Does b change? **yes** Before swap: $b = r\frac{s_y}{s_x}$ After swap: $b = r\frac{s_x}{s_y}$

original

$b_1 = r\frac{S_y}{S_x}\ \left(\frac{\text{steps}}{\text{in}}\right)$

$-.5728 = (-.8734)\frac{S_y}{S_x}$

so $\frac{S_y}{S_x} = \frac{-.5728}{-.8734}$

swapped

$b_2 = r\frac{S_x}{S_y}\ \left(\frac{\text{in}}{\text{step}}\right)$

$b_2 = (-.8734)\left(\frac{-.8734}{-.5728}\right)$

$b_2 = -1.332\ \frac{\text{in}}{\text{step}}$

$\widehat{height} = \boxed{a} - 1.332(\text{steps})$

$(\overline{steps}, \overline{height})$ is on LSRL

$(16.1389, 65.8333)$

$(65.8333) = a - 1.332(16.1389)$

$a = 87.33$

$\boxed{\widehat{height} = 87.33 - 1.332(\text{steps})}$

3) Given data value std dev in x, find corresponding std dev in y

If a student's height is 1.2 standard deviations above the mean height of the class, how many standard deviations above the mean number of steps would you predict this student's number of steps to be?

(When considering standard deviations, these are z-scores, so $s_y = s_x = 1$)

$$\text{slope} = \frac{z_y}{z_x} = b = r\frac{s_y}{s_x}$$

$$\frac{z_y}{z_x} = (-.8734)\frac{(1)}{(1)}$$

$$z_y = (-.8734)z_x$$

$$z_y = (-.8734)1.2$$

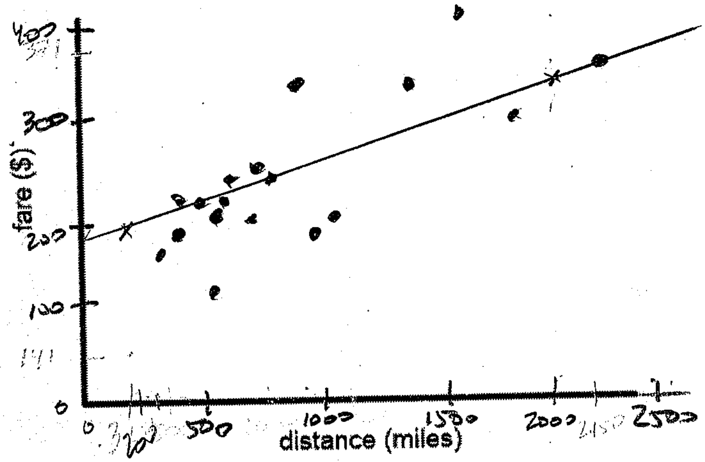
$$\boxed{z_y = -1.048}$$

This student is predicted to have # steps 1.048 standard deviation below the mean number of steps.

Flights from Atlanta - The following data explore the relationship between the cost of an airplane ticket (fare) and the distance flown (from Atlanta to various cities):

Atlanta to:	Distance	Fare
Baltimore	568	219
Boston	933	222
Dallas	720	249
Denver	1190	308
Detroit	602	249
Kansas City	683	141
Las Vegas	1719	252
Miami	589	229
Memphis	327	183
Minneapolis	894	209
New Orleans	419	199
NY	749	248
Okla City	749	301
Orlando	392	238
Philadelphia	657	205
St Louis	461	232
Salt Lake	1565	371
Seattle	2150	343

#17. Enter the data in your calculator and sketch the scatter plot (include value labels for the axes, but do not sketch in an LSRL):



#18. Use your calculator's LinReg feature to find the LSRL: $\hat{y} = 177.215 + 0.0786x$

x: distance (miles)
y: air-fare ($)

#19. Find r^2 and r. $r^2 = .4820$
$r = .6943$

#20. Explain what r^2 means in this context.

About 48.2% of the variation in air-fares is explained by the LSRL model which relates air-fare to distance.

#21. Explain what the slope of the LSRL means in this context.

For every 1 additional mile of distance, air-fare increases by $0.0786, on average.

#22. Explain what the y-intercept means in this context.

For a flight of 0 miles, the predicted air fare is $177.215.
(or this is the 'fixed cost' always present, regardless of distance)

#23. Estimate the air fares for a 200-mile flight and for a 2000-mile flight.

$\hat{y} = 177.21 + 0.0786(200) = \boxed{\$192.92}$

$\hat{y} = 177.21 + 0.0786(2000) = \boxed{\$334.42}$

#24. Using the estimates from #23, and #24, draw the LSRL on the scatter plot in #17.

#25. The fare to fly to Los Angeles, 1719 miles from Atlanta, is $212. Find the residual for this flight.

predicted $\widehat{\text{fare}} = 177.21 + 0.0786(1719) = \$312.31$

actual fare = $212

resid = e = actual − predicted = $212 - 312.31 = \boxed{-\$100.31}$

#26. What does a positive residual mean in this problem's context?

Actual air fare is higher than the model predicts for that distance.

#27. What does a negative residual mean in this problem's context?

Actual air fare is lower than the model predicts for that distance.

#28. What if we had not given you the data set, but instead only the summary statistics:

Determine the LSRL from the summary statistics (show your work).

Atlanta to:	Distance x	Fare y
Summary Statistics		
Mean	853.7	244.33
St Dev	497.8	56.37
Correlation	0.694	

$b = r\frac{s_y}{s_x}$

$b = (0.694)\frac{(56.37)}{(497.8)}$

$b = .0786$

$\hat{y} = a + 0.0786x$

$(853.7, 244.33)$ is on LSRL

$(244.33) = a + 0.0786(853.7)$

$a = 177.23$

$\hat{y} = 177.23 + 0.0786x$

x: distance (miles)

y: air fare ($)

#29. What if no data or statistics were given, but you had the output of a software package?

How would you determine the LSRL and r^2 from this software output?

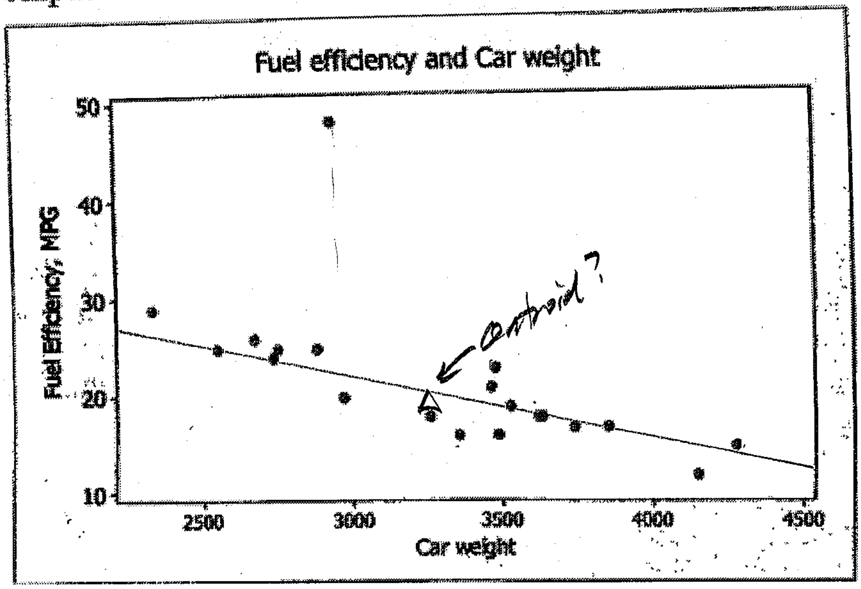
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.21452   19.99315   8.864 1.43e-07 ***
distance      0.07862    0.02037   3.859  0.00139 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `

Residual standard error: 41.82 on 16 degrees of freedom
Multiple R-squared:  0.482,     Adjusted R-squared:  0.4496
F-statistic: 14.89 on 1 and 16 DF,  p-value: 0.00139
```

$\hat{y} = 177.21452 + 0.07862x$

x: distance (miles)

y: air fare ($)

The scatterplot below shows the fuel efficiency (in miles per gallon) and weight (in pounds) of twenty 2009 subcompact cars.



#30. There is one "unusual point" on the graph. Explain what is "unusual" about this car.

The point ~(2900, 48) has an unusually high fuel efficiency for that car's weight (much higher than a linear model would predict) Perhaps it is a hybrid-electric car?

#31. What effect would removing the "unusual point" have on the slope of the LSRL for this data? Justify your answer.

The point has a fairly small, but not zero, leverage (to the left of $(\bar{x}, \bar{y})$) so it will have some effect on slope, It is currently pulling up on the left side so removing this point would cause the slope to increase (become less negative), at least slightly.

#32. What effect would removing the "unusual point" have on the correlation? Justify your answer.

Removing this point means the points are now closer on average to the LSRL so the correlation would strengthen.

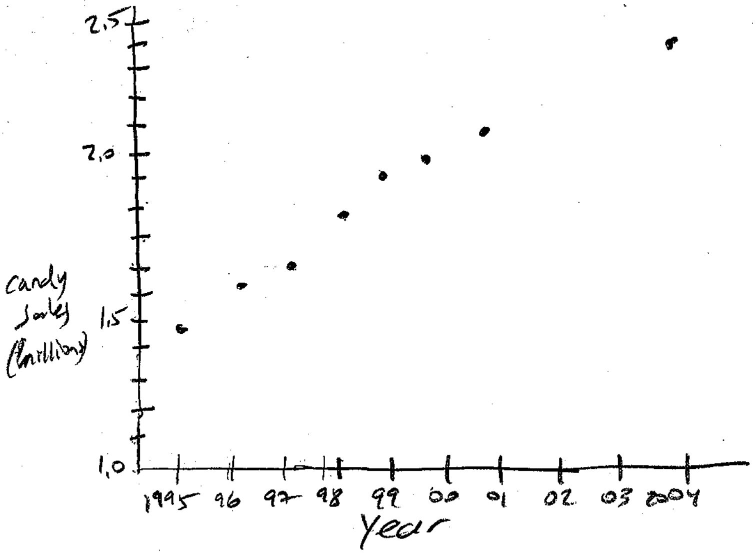
Here, since r is negative, r would become more negative.

#33. If we converted the car weights to metric tons (1 metric ton = 2205 pounds), how would the correlation change? Explain.

r would not change.

Halloween Candy Sales

Year	Candy Sales (millions of dollars)
1995	1.474
1996	1.66
1997	1.708
1998	1.787
1999	1.896
2000	1.985
2001	2.035
2004	2.41



#34. How would you describe the association between the year and candy sales?

There is a linear, strong, positive association between year and candy sales.

#35. What is the explanatory variable? year

What is the response variable? candy sales

#36. Fit a LSRL to the data and write it down. $\hat{y} = -191.783 + 0.0969x$

x: year
y: candy sales ($ millions)

#37. Predict candy sales for 2003.

$\hat{y} = -191.783 + 0.0969(2003) = \boxed{\$2.3077 \text{ million}}$

#38. Predict candy sales for 2013.

$\hat{y} = -191.783 + 0.0969(2013) = \boxed{\$3.2767 \text{ million}}$

#39. What might cause you not to report the value you calculated in #38?

2013 is ~~far~~ beyond years where we have data (1995–2004)
this is extrapolation.

#40. What is the residual for 1998?

$\hat{y} = -191.783 + 0.0969(1998) = 1.8232$ resid $= y - \hat{y}$

$y = 1.787$

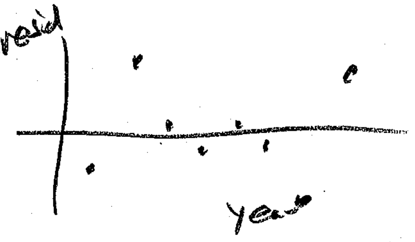
$= 1.787 - 1.8232$

$= -\$0.0362$ million

$(-\$36,200)$

#41. Roughly sketch and describe the residual plot. Is a linear model appropriate for this data?

(use calculator scatterplot L1, RESID)



Yes, because there is no pattern in the residuals

#42. What is r? .9927 What does it measure? Strength of association, There is a strong, positive, linear association between candy sales and year.

#43. What is r^2? .9855 Interpret r^2 in the context of the problem.
About 98.55% of the variation in candy sales is explained by the LSRL model which relates candy sales to year.

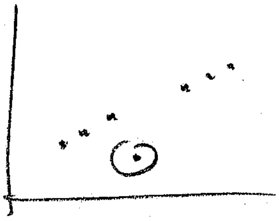
#44. What is the slope, b, of the LSRL? .0969 Interpret b in the context of the problem.
$million/yr

For every additional year, candy sales increase by $.0969 million ($96900), on average.

#45. Change (1998, 1.787) to (1998, 1.387). How does this affect…

…..b? increases from .0969 to .1019

……r? decreases from .9927 to .8944



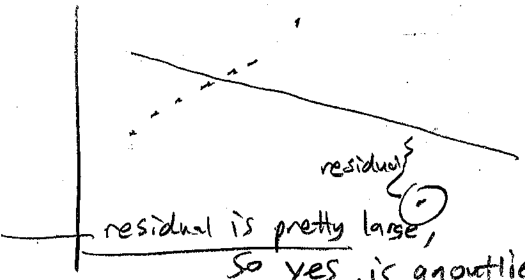
Is this point an outlier? (Explain)
yes (it is separated from the group)

Is this an influential point? (Explain)
no (the effect on the slope was small), This is because the point does not have much leverage (fairly close to centroid horizontally

#46. Change (1998, 1.387) to (1998, 1.787). Add the point (2013, 0.956) How does this affect…

…..b? decreases from 0.0969 to − 0.0242

……r? decrease from 0.9927 to − 0.3276



residual is pretty large, So yes, is an outlier

Is this point an outlier? (Explain)
It is definitely an unusual point. For it to be an "outlier" it has to have a large residual!

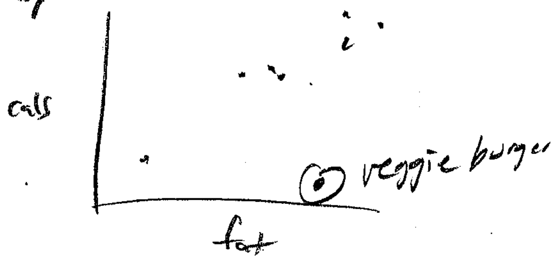
Is this an influential point? (Explain)
yes, adding the point dramatically changes the slope (and actually makes it negative) (influential on the slope and the y-intercep

Burgers The following data were measured for a sample of burgers:

x **Fat**	19	31	34	35	39	39	43
y **Calories**	410	580	590	570	640	680	660

#47. Describe the association of calorie content vs. fat.

there is a fairly strong, positive, linear relationship between calories and fat.



#48. Explanatory variable: fat

Response variable: calories

#49. Find the LSRL: $\hat{y} = 210.954 + 11.056x$

x: fat (g)

y: calories

#50. Predict the calories for a burger with 30 grams of fat.

$\hat{y} = 210.954 + 11.056(30) = \boxed{542.6 \text{ calories}}$

#51. Find r and r^2. Interpret r^2 in the context of the problem.

$r = .9606$, $r^2 = .9228$

About 92% of the variation in calories is explained by the LSRL model which relates calories to fat.

#52. Find the slope, and interpret the slope in the context of the problem.

$b = 11.056$ cals/g

For every 1 additional gram of fat, calories increase by 11.056, on average.

#53. Add in a veggie burger to the data that has 36 grams of fat and 120 calories.

Is this an outlier? yes (has a large residual)

Is this an influential (on slope) point? Not really, it has low leverage, so only a small effect on slope.

What effect does adding this veggie burger have on slope?

Adding the veggie burger decreases the slope from 11.056 to 9.0625.

What effect does adding this veggie burger have on correlation?

Decreases the correlation (r from 0.9606 to 0.3536)

What is the residual for the new data point? (should redo the LSRL w/ point include first):

$\hat{y} = 218.594 + 9.0625x$ $\hat{y} = 218.594 + 9.0625(36) = 544.844$

$y = 120$

resid $= 120 - 544.844 = \boxed{-424.8 \text{ calories}}$

Chapter 7 Practice Quiz

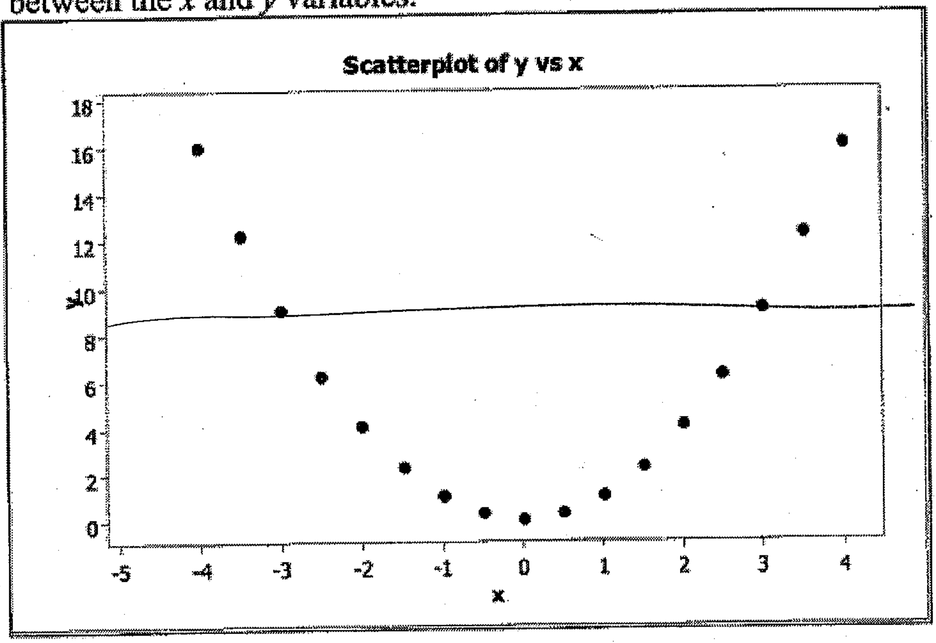
1. After conducting a survey of his students, a professor reported that "There appears to be a strong correlation between grade point average and whether or not a student works." Comment on this observation.

(categorical)

"Correlation" refers specifically to r, the measure of the strength of the association. You can't compute r unless both variables are numerical.

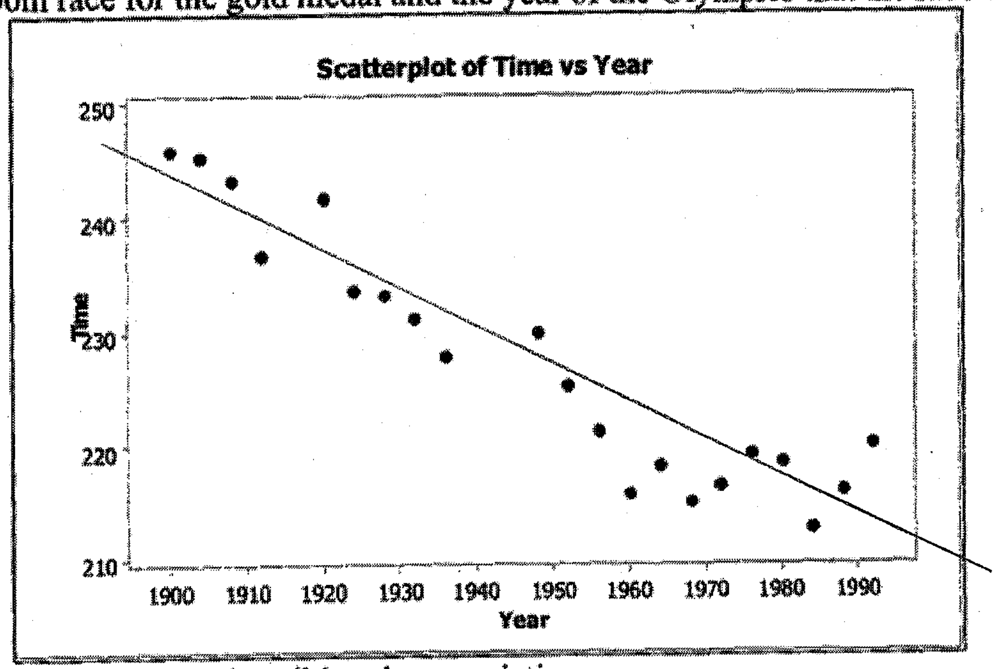
(You could use the word "association" and divide GPA in categories, then use segmented bargraphs to assess)

2. The following scatterplot shows a relationship between x and y that results in a correlation coefficient of $r = 0$. Explain why $r = 0$ in this situation even though there appears to be a strong relationship between the x and y variables.



r measures the strength of linear associations. Although this data shows a strong association, it is not a linear association.

3. The following scatterplot shows the relationship between the time (in seconds) it took men to run the 1500m race for the gold medal and the year of the Olympics that the race was run in:



a. Write a few sentences describing the association.

There is a fairly strong, negative, linear association between race time and year.

b. Estimate the correlation. $r = -0.94$

(any vale from −0.7 to −0.98 is okay)

4. Identify what is wrong with each of the following statements:

a. The correlation between Olympic gold medal times for the 800m hurdles and year is –0.66 seconds per year. r has no units

b. The correlation between Olympic gold medal times for the 100m dash and year is -1.37.

range for r is −1 to 1 (out of range)

c. Since the correlation between Olympic gold medal times for the 800m hurdles and 100m dash is –0. 41, the correlation between times for the 100m dash and the 800m hurdles is +0.41. r is still −0.41 (not affected by swapping variables)

(slope would change, but would need to use $b = r\frac{s_y}{s_x}$ to find it)

d. If we were to measure Olympic gold medal times for the 800m hurdles in minutes instead of seconds, the correlation would be –0.66/60 = –0.011.

No, r would still be −0.66

(units don't affect r which uses z-scores)

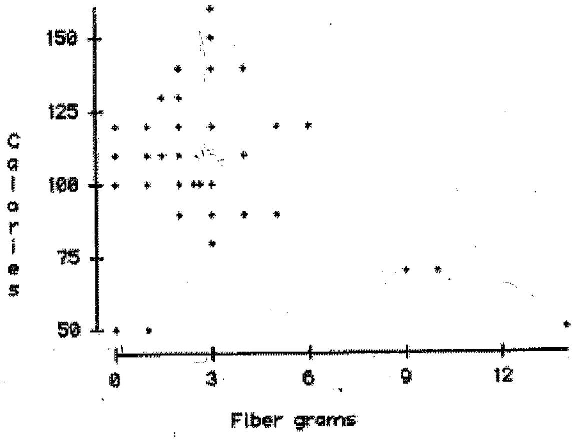
$$r = \sum \frac{z_x z_y}{n-1}$$

Chapter 9 Practice Quiz

Current research states that a good diet should contain 20-35 grams of dietary fiber. Research also states that each day should start with a healthy breakfast. The nutritional information for 77 breakfast cereals was reviewed to find the grams of fiber and the number of calories per serving. The scatterplot below shows the relationship between fiber and calories for the cereals.

1. Do you think there is a clear pattern? Describe the association between fiber and calories.

There is a weak, linear, negative association between fiber and calories.



2. Comment on any unusual data point or points in the data set. Explain.

The 3 lower right points are unusual with high leverage and will have a large effect on the LSRL slope. (making the association negative). There are also 2 lower left outliers which are trying to make the association positive. But these points have less leverage than the lower right outliers and have less influence on the slope.

3. Do you think a model could accurately predict the number of calories in a serving of cereal that has 22 grams of fiber? Explain.

No, this would be extrapolation. (If you account for axes zeros and extend the graph, the model would probably predict negative calories for 22g fiber.)